Open

The ISME Journal (2011) 5, 1565–1567

www.nature.com/ismej

COMMENTARY

The genomic standards consortium: bringing standards to life for microbial ecology

Pelin Yilmaz, Jack A Gilbert, Rob Knight, Linda Amaral-Zettler, Ilene Karsch-Mizrachi, Guy Cochrane, Yasukazu Nakamura, Susanna-Assunta Sansone, Frank Oliver Glöckner and Dawn Field

The ISME Journal (2011) **5,** 1565–1567; doi:10.1038/ismej.2011.39; published online 7 April 2011

Adoption of easy-to-follow standards will vastly improve our ability to interpret data from genomes, metagenomes and marker studies

Interest in sampling of diverse environments, combined with advances in high-throughput sequencing, vastly accelerates the pace at which new genomes and metagenomes are generated. For example, as of January 2011, 12 500 user-generated metagenomes have been submitted to the public MG-RAST Annotation server (http://metagenomics.nmpdr.org; Meyer *et al.*, 2008), >90% of which were produced using high-throughput sequencing methodologies. We have entered into an era of 'mega-sequencing projects' that include the Genomic Encyclopaedia of Bacteria and Archaea project (http://www.jgi.doe.gov/programs/GEBA), the Microbial Earth Project (http://genome.jgi-psf.org/programs/bacteria-archaea/MEP/index.jsf), the Human Microbiome Project (http://nihroadmap.nih.gov/hmp), the Metagenomics of the Human Intestinal Tract consortium (http://www.metahit.eu), the Terragenome Initiative (http://www.terragenome.org), the Tara Oceans Expedition (http://oceans.taraexpeditions.org), the National Ecological Observatory Network (NEON-http://www.neoninc.org), the International Census of Marine Microbes (ICoMM-http://icomm.mbl.edu), Microbial Inventory Research Across Diverse Aquatic Long-Term Ecological Research Sites (http://amarallab.mbl.edu/mirada/mirada.html), the Earth Microbiome Project (http://www.earthmicrobiome.org) and other funded and unfunded projects, with many more visionary projects on the horizon.

Additionally, studies of emerging metatranscriptomes (community transcript profiles), metaproteomes (community protein profiles) and metametabolomes (community metabolite profiles) now complement genomes and metagenomes. Comparative studies of multi-omic data sets from the same community hold the promise of unparalleled insights into fundamental questions across a range of fields including evolution, ecology, environmental science, physiology and medicine. Advances stem from improvements in the annotation and quantification of genes, pathways, organisms and consortia within these communities. We are just starting to exploit these technologies to understand the microbial world, and have only scratched the surface in terms of sampling microbial diversity across temporal and spatial scales (Delmotte *et al.*, 2009; Gilbert *et al.*, 2010a). To fully exploit the promise of these data, we need both scientific innovation and community agreement on how to provide appropriate stewardship of these resources for the benefit of all.

Although we have collected billions of nucleic-acid sequences from thousands of ecosystems, illuminating uncharacterized microbial lifestyles remains far from trivial. For example, in each analysed genome or metagenome, about 40% of the putative protein-coding genes cannot be assigned to any known function or taxon. Only 42% of the 61 known bacterial phyla have even a single cultured representative (Hugenholtz and Kyrpides, 2009), with the remainder being known only from 16S rRNA gene environmental surveys. Surprisingly, only 14% of cultured bacterial taxa have a single complete genome sequenced. Holistic approaches that will centralize (meta) omics data are needed, which will allow investigators to analyze these data within the context of space, time, habitat and characteristics of the environment. Networks of information arising from these studies will allow us to describe and predict ecological patterns of organisms, genes, transcripts and proteins.

One key insight into the function of a gene or organism is the environment where it occurs. Collection of contextual (meta) data, which delineates the source of a sequence in terms of the space, time, habitat and characteristics of the environment, is thus essential in interpreting these unknown genes and species, as well as gaining new insights into the known fraction. Although early comparative studies of metagenomes (Tringe *et al.*, 2005) relied on a few, deeply sequenced samples, the experience

from 16S rRNA gene surveys suggests that additional insight is gained from observing spatial and temporal variation across hundreds of samples, whether examining the distribution of bacteria in soils across a continent (Lauber *et al.*, 2009) or various skin sites from many subjects (Grice *et al.*, 2009).

At present, the valuable contextual data halo is often missing for sequences deposited in the International Nucleotide Sequence Database Collaboration (INSDC; GenBank, European Nucleotide Archive (ENA, including EMBL-Bank) and the DNA Databank of Japan (DDBJ)). This leaves researchers in the position of searching in electronic resources, literature or contacting the authors for even the most basic contextual data, such as geographic location, date and time of sampling or the habitat where the sample was obtained. Molecular ecologists should immediately recognize the inherent value of these data to the community, because without them their own sequence data sets will have extremely limited comparability with the wealth of other data available. Sequences without contextual data are like unlabeled cans in a supermarket—you do not know what you are purchasing until you open it and examine the contents. The present inability to automatically retrieve rich contextual data hampers comparative research, and constitutes a considerable misuse of the vast global resources currently being applied to microbial ecology. Just as food-safety laws emphasize clear and accurate labeling based on the product, process and producer, so should sequence data be properly annotated.

Standardization of the required information will greatly facilitate the annotation of sequence data. To achieve this, we must first have community collaboration and participation. Second, as a result of this collaboration, a contextual data set must be standardized in terms of content, syntax and terminology to which the community can adhere. In 2005, members of the community came together to form the Genomic Standards Consortium (GSC), an open-membership working body with the stated mission of working towards better descriptions of our genomes, metagenomes and related data (http://www.gensc.org). Supported by the expertize of the members involved in many of the aforementioned mega-sequencing projects, the GSC has formalized contextual data requirements for genomes and metagenomes as the Minimum Information about a Genome/Metagenome Sequence checklist (MIGS/MIMS) (Field *et al.*, 2008). Furthermore, to cover the description of phylogenetic and functional marker genes an extended standard, the Minimum Information about a MARKer gene Sequence (MIMARKS) checklist (http://gensc.org/gc_wiki/index.php/MIMARKS) has been developed (Yilmaz *et al.*, 2011). This family of minimum information checklists provides researchers with a condensed set of contextual data requirements, which range from description of the environment to sampling and sequencing procedures. The GSC is also driving the evolution of omics data sharing in a broader context through participation in the BioSharing (http://biosharing.org) portal. This forum aims to enable a broader dialog among funders, journals, standards and technology developers, and researchers on the critical issue of data sharing within the metagenomics community and beyond (Field *et al.*, 2009). It provides an example of what an infrastructure to support standards-compliant reporting of contextual data might look like; as well as encouraging and enabling curation at community level (Rocca-Serra *et al.*, 2010; http://isatab.sourceforge.net).

The primary sequence databases' adoption of these standards is integral to their success. The INSDC partners have recognized this support for submission of compliant data sets with the adoption of an official keyword for the family of minimum standards reserved for compliant INSDC sequence records. Additionally, the development of a number of tools and formats to aid in data exchange (Kottmann *et al.*, 2008) and compliance during sequence submissions with these standards is ongoing within specialized genomics and metagenomics resources.

The application of high-throughput sequencing technologies has transformed the way microbial ecologists approach questions in their field (Gilbert *et al.*, 2010b). The shift of sequencing capacity to individual labs is creating a data bonanza. With appropriate contextual information, these data sets could herald a new era of discovery for microbial ecology. This will only be possible, if each study, from each environment, and from each lab maintains, at the very least, a minimum contextual data standard to facilitate cross-comparison and meta-analysis of global microbial communities. Inadequate implementation of these standards threatens progress in our field of research, as we will lose the best opportunity to produce a complete mechanistic understanding of microbial life. Every investigator will benefit immensely by being able to obtain a rapid, comprehensive answer to the question 'Have my microbes been seen before, and, if so, where, with whom, and what were they doing? Only by accepting the relatively small responsibility of entering their own contextual data into a global system will they realize this dream. Just as standardized deposition of sequence data contributed an immensely valuable resource, standardization of contextual data will allow us to reap vast dividends for decades to come and enable us to finally escape the burden of 'my sequence matches 1500 uncultured environmental isolates—now what'?

To provide a better understanding of the requirements, we included three examples for MIGS, MIMS and MIMARKS compliant data sets in the Supplementary Table 1. Supplementary File 2 provides links to detailed submission and compliance guidelines.

With this open letter to the ISME community, we not only hope to advertise the existence of the GSC and invite more microbial ecologists investigating marker genes and doing 'omics' work to join us, but also make a call for compliance with current and future GSC standards. To learn how to describe your data according to MIGS/MIMS/MIMARKS (MIxS) standards, please visit the GSC website for details and options for submitting compliant data sets into public domain databases (http://gensc.org/gc_wiki/index.php/MIGS/MIMS/MIMARKS).

P Yilmaz is at Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
P Yilmaz is at Jacobs University Bremen gGmbH, Bremen, Germany;
JA Gilbert is at Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
JA Gilbert is at Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA;
R Knight is at Howard Hughes Medical Institute and Department of Chemistry & Biochemistry, University of Colorado at Boulder, Boulder, CO, USA;
L Amaral-Zettler is at Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, MA, USA;
I Karsch-Mizrachi is at National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA;
G Cochrane is at EMBL Outstation, The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK;
Y Nakamura is at Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima, Japan;
S-A Sansone is at Oxford e-Research Centre, University of Oxford, Oxford, UK;
FO Glöckner is at Microbial Genomics and Bioinformatics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany
FO Glöckner is at Jacobs University Bremen gGmbH, Bremen, Germany and
D Field is at NERC Centre for Ecology and Hydrology, Oxford, UK
E-mail: fog@mpi-bremen.de

References

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R *et al.* (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc Natl Acad Sci USA* **106**: 16428–16433.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P *et al.* (2008). The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**: 541–547.

Field D, Sansone S-A, Collis A, Booth T, Dukes P, Gregurick SK *et al.* (2009). 'Omics data sharing'. *Science* **326**: 234–236.

Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B *et al.* (2010a). The taxonomic and functional diversity of microbes at a temperate coastal site: A 'multi-omic' study of seasonal and diel temporal variation. *PLoS One* **5**: e15545.

Gilbert JA, Meyer F, Bailey MJ. (2010b). The Future of microbial metagenomics (or is ignorance bliss?). *ISME J*; e-pub ahead of print 25 November 2010, doi:10.1038/ismej.2010.178.

Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC *et al.* (2009). Topographical and temporal diversity of the human skin microbiome. *Science* **324**: 1190–1192.

Hugenholtz P, Kyrpides NC. (2009). A changing of the guard. *Environ Microbiol* **11**: 551–553.

Kottmann R, Gray T, Murphy S, Kagan L, Kravitz S, Lombardot T *et al.* (2008). A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS* **12**: 115–121.

Lauber CL, Hamady M, Knight R, Fierer N. (2009). Soil pH as a predictor of soil bacterial community structure at the continental scale: A pyrosequencing-based assessment. *Appl Environ Microbiol* **75**: 5111–5120.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform* **9**: 386.

Rocca-Serra P, Brandizi M, Maguire E, Sklyar N, Taylor C, Begley K *et al.* (2010). ISA infrastructure: supporting standards-compliant experimental reporting and enabling curation at the community level. *Bioinformatics* **26**: 2354–2356.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.

Yilmaz P, Kottman R, Field D, Knight R, Cole JR, Amaral-Zettler L *et al.* (2011). The minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specification. *Nat Biotechnol*; accepted 18 February 2011.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)